Energy Proceedings
Vol 55, 2025
ISSN 2004-2965

Exploring the Effectiveness of SHAP for Feature Selection in National Demand Forecasting#

Corné van Zyl*, Xianming Ye, Raj Naidoo

Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria, 0002, South Africa

*Corresponding Author: c.vanzyl@tuks.co.za

ABSTRACT

This study evaluates the effectiveness of SHAP (SHapley Additive exPlanations) as a feature selection tool for national demand forecasting, addressing the need for a comprehensive understanding of its strengths and limitations. We benchmark SHAP against common selection methods, including variance-based selection, XGBoost, and NSGA-II. Using a 1D Convolutional Neural Network to forecast hourly national energy demand in Panama, we compare the performance of features selected by each method. Our results indicate that SHAP is generally outperformed by XGBoost and may not be the most effective initial approach for feature selection, nor does it consistently rank features by their predictive value. However, SHAP proves more effective in refining smaller sets of features, maintaining model accuracy by eliminating non-contributive features.

Keywords: feature selection, SHAP, national demand forecasting, explainable artificial intelligence

NONMENCLATURE

Abbreviations	
CNN	Convolutional Neural Network
XAI	Explainable Artificial Intelligence
SHAP	Shapley Additive exPlanations

1. INTRODUCTION

Accurate demand forecasting is essential for efficient resource management and planning but is complicated by various factors such as weather conditions, electricity prices, and fluctuating consumer behaviour. Data-driven modelling is often used for this purpose, however, these models are sensitive to the quantity and type of training data. Adding too much information to the model can unintentionally include irrelevant or redundant variables, leading to long training times and overfitting, where the model learns noise instead of the underlying patterns. Thus, selecting the right features is crucial for optimal model performance. Feature selection reduces data dimensionality, decreases training time, and simplifies model complexity, often improving forecast accuracy. Moreover, models with fewer features tend to be more interpretable [1].

Despite decades of research, feature selection remains challenging, with no single method suited for all datasets. Filter methods, which evaluate features based on intrinsic data characteristics without a learning algorithm, are computationally efficient and ideal for large datasets. However, they often miss complex feature interactions, leading to suboptimal feature selection [1]. In contrast, wrapper methods assess features based on their impact on model performance, optimizing for accuracy. While this approach generally yields highest accuracy, it incurs substantial computational costs due to iterative model training, making it impractical for high-dimensional datasets[2]. Embedded methods integrate feature selection within the learning algorithm, balancing accuracy and efficiency by directly linking feature importance to model training. However, embedded methods are limited to specific algorithms, and the selected features may not perform well in different models [1].

In the emerging domain of Explainable Artificial Intelligence (XAI), XAI techniques are also applied for feature selection. The core rationale for using XAI in this context is its ability to extract the abstract learned patterns from the model and identify and quantify the importance of each input feature to the model's predictions through feature attributions. These attributions are scores that indicate the extent to which each feature contributes to the model's output. By retraining the model using only the features that significantly influence its predictions, features with low attribution scores can be eliminated without compromising the model's accuracy.

This is a paper for the 16th International Conference on Applied Energy (ICAE2024), Sep. 1-5, 2024, Niigata, Japan.

XAI methods present a compelling alternative to traditional feature selection techniques. Unlike computationally intensive wrapper methods, XAI methods can determine feature importance using the model's learned patterns, significantly reducing computation time. Furthermore, since XAI methods extract complex feature interactions learned by the model, they have been shown to consistently outperform filter methods [3][4]. XAI methods are also applied directly to the desired forecasting model, avoiding the need to train alternative models that might not effectively capture the relevant patterns for the specific forecasting model. In scenarios such as national demand forecasting, where datasets are often characterised by high dimensionality, the direct application of XAI methods for feature selection has demonstrated potential [5].

This study focuses on SHAP (Shapley Additive exPlanations) [6], a prominent XAI method recognised for its strong theoretical foundation in cooperative game theory and adaptability to any machine learning model. While SHAP has proven to be effective for feature selection [5][7][8], its performance in time series feature selection and its comparison to meta-heuristic based wrapper methods, which typically yield the highest model accuracy compared to greed search strategies, remain underexplored. Given the high computational demand of optimal feature selection, a method that approximates the best solution with fewer resources is appealing. If SHAP falls short of selecting features that produce highly accurate models, it might justify the use of more resource-intensive methods or prompt the exploration of alternatives. This underscores the importance of rigorously evaluating SHAP to ensure strategic and justified allocation of computational resources in feature selection.

In light of the growing prevalent use of SHAP as a feature selection tool, this study critically evaluates SHAP's effectiveness as a feature selection tool for national demand forecasting, addressing the need for a comprehensive understanding of its strengths and limitations. We benchmark SHAP against optimal feature combinations, evaluating its overall effectiveness and other common selection methods, providing valuable insights for researchers designing forecasting models. Our novel contributions include: (1) the first evaluation of SHAP for feature selection in national demand forecasting, (2) pioneering the use of 1D CNNs for efficient wrapper-based time series feature selection, and (3) assessing SHAP's ability to identify both significant and non-contributing features. This research enables more informed application of SHAP in interpretable machine learning and demand forecasting, enhancing model accuracy and interpretability in these critical domains.

2. METHODOLOGY

2.1 Data processing

The dataset is augmented to enhance representation and increase dimensionality by separating date and time into distinct components like hour, day, and month, and marking holidays and weekends with binary indicators. Weather-related features are added with a 24-hour lead time to simulate forecasts, alongside similar series for holidays and weekends to model upcoming events. Further augmentation includes adding transformations such as sinusoidal modifications of periodic features to capture cyclical patterns, and differencing all original features to introduce redundancy and aid in feature analysis.

For model training and testing, a sliding window algorithm is used to create a batched time series dataset. The window size is determined by partial auto-correlation analysis and advances one step at a time. Finally, data is vectorized to meet the tabular format required by SHAP for analysis

2.2 Forecasting model - 1D CNN

Our study uses a 1D Convolutional Neural Network (CNN) to model multivariate time series national demand data. The architecture begins with a vector input layer, reshaped into a 2D array where each feature's time series aligns with a CNN channel. The model includes a single convolutional layer with filters applied across the temporal dimension, a max-pooling layer to reduce temporal dimensionality, and fully connected layers for prediction. The output layer's neurons correspond to the required hourly predictions

2.3 Feature selection methods

2.3.1 Variance Method

The variance method for feature selection evaluates features based on their variance within the dataset. Features with low variance are considered less informative and are removed, while high-variance features are retained, as they are likely to provide more significant information for the model

2.3.2 Extreme Gradient Boosting

Feature selection in XGBoost is based on the importance scores assigned to each feature during

model training, with gain being a primary metric. Gain measures the improvement in accuracy brought by a feature to the branches it is on. Features are ranked based on their total gain, and the least important ones can be removed to enhance model efficiency and reduce overfitting.

2.3.3 SHAP

This study leverages the SHapley Additive exPlanations (SHAP) algorithm to quantify the contribution of each feature to model predictions. By aggregating SHAP values across multiple predictions, we can determine the average contribution of a feature.

2.3.4 NSGA-II

Non-dominated Sorting Genetic Algorithm II (NSGA-II) is a multi-objective optimization algorithm used to select the optimal subsets of features and minimise both model error (MSE) and feature count. It evolves a population of solutions over generations, employing non-dominated sorting and crowding distance to maintain diversity [9]. The algorithm is implemented using the DEAP (Distributed Evolutionary Algorithms in Python) framework [10]

3. CASE STUDY

This study focuses on forecasting the hourly national energy demand in Panama, a tropical country in Central America with minor seasonal changes in temperature. The dataset includes weather conditions from three major cities: Tocumen (near the capital, Panama City), David City, and Santiago, and information on public and school holidays. The data covers the period from 3 January 2015 to 27 June 2020, but data from 2020 is excluded due to the COVID-19 pandemic's impact on energy demand.

Preprocessing and augmenting the original dataset produces approximately 43,200 data points for training and testing, comprising 65 input features detailed in Table 1. The 1D CNN is trained on data from 2015 to the end of 2018. The model inputs consist of a vectorised format of a 168-hour look-back window of multivariate hourly data with 65 features, used to predict energy demand for the next 24 hours. Testing is conducted using 2019 data, with each test starting at 00:00 to forecast demand for the following 24 hours, resulting in 356 non-overlapping test cases throughout the year.

The experiments are carried out using Python 3.10, Keras 2.8, TensorFlow 2.8.0, and SHAP 0.41.0, run on an Intel i9-12900 CPU with 32 GB RAM. To expedite the NSGA-II feature selection method, distributive computing is employed using 10 computers, each running 3 instances of the model training script simultaneously, where models are trained using features selected by the GA.

Table 1: Description of dataset features.

Description	Feature name	Unit/Range
Demand	Demand	MW
Weather features		
Temperature	Temp - Toc	˚C
Relative humidity	Hum - Toc	%
Precipitation	Rain - Toc	$\mathrm{liters/m^2}$
Wind speed	Wind - Toc	$\mathrm{m/s}$
Calendar features		
Holiday ID	Hol ID	[1-22]
Holiday indicator	is hol	[0]
School holiday	is scho	[0]
Weekend indicator	isweekend	[0]
Date and time features		
Hour of day	hour	[0-23]
Day of the week	dayofweek	[0-6] Monday = 0
Day of month	day	[1-31]
Month of the year	month	[1-12]
Year	year	[2015-2019]
Augmented features		
Lead features	Future - [Weather]	
Differenced features	diff - [Demand]	
Sinusoidal features	cyclic - [Hol ID]	

Toc = Tocumen, San = Santiago City, Dav = David City

3.1 Computation time

Table 2 shows the computational time for each feature selection method. For SHAP and XGBoost, the times include both model training and feature importance calculation. The variance method is significantly faster due to its simple calculations and lack of model training. In contrast, SHAP and XGBoost require model training before calculating feature importance. For XGBoost, the reported time is mainly for model training, with feature importance derived directly from the model. SHAP is more time-intensive due to the demanding SHAP value calculations. The NSGA-II method is the most time-consuming, taking roughly 3 days to complete with the aid of distributive computing.

Table 2: Performance times for different feature selection methods.

Feature selection method	Time
Variance method	32 ms
XGBoost	8.08 min
SHAP	14.9 min
NSGA-II	40.67 days

3.2 Performance of selected features

The performance of the features selected by the various methods are presented in Fig. 1.

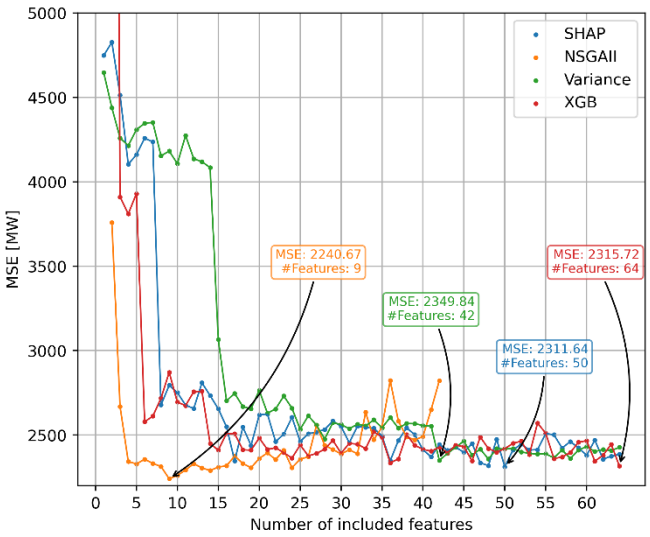

Fig. 1 Comparison of the minimum MSE for selected number of features.

Our study found no significant difference between SHAP, XGBoost, and the Variance method's selected features when using the top 42 to 64 ranked features of each method, but differences emerged at 20 features and below. The Variance method resulted in the poorest model performance, while XGBoost selected features generally outperformed those selected by SHAP. XGBoost also retained features with significant predictive content better than SHAP, particularly at 6 and 14 features.

SHAP ranked 'Demand' and 'diff demand' as the two most significant features indicating a strong autoregressive component in the model but this also results in redundancy in selected features. A similar redundancy was observed with XGBoost combining 'hour' and 'cyclic hour' as top features.

While no such claim has been explicitly made, it is clear when compared to the optimally selected features, that SHAP does not rank features by their contribution to forecasting accuracy or quantify the predictive content of features. However, the ambiguous language used to describe SHAP values and "feature importance" in general as "a measure of a feature's significance to the model", could be misinterpreted as a measure of predictive content. This could lead to the incorrect belief that the more "important" or "significant" a feature is, the more justification there is to include it in the model.

3.3 SHAP for refined feature selection

One claim that may hold merit is that features with low or zero attribution scores can be removed without adversely affecting model accuracy. While this appears true for up to 40 top-ranked features, the comparable results between SHAP and Variance methods in this range do not strongly support this claim. To evaluate SHAP's effectiveness in refining feature selection, we conducted an experiment using the top 16 features identified by NSGA-II. Since the optimal number of features is determined to be 9, selecting 16 features allowed us to investigate the contribution of additional features that might not be essential. We trained a 1D CNN with these features, calculated average SHAP values, and excluded those with zero attribution scores, resulting in 11 features. Table 3 presents the mean and standard deviation of performance metrics across 50 experimental runs using this refined feature set, demonstrating the impact of SHAP-based feature selection on model accuracy and consistency.

Table 3: Comparative performance metrics of CNN with refined feature set based on non-zero average SHAP values.

Method	Num Features	MSE	R^2
NSGA-II+SHAP	11	2365.87 ± 103.29	0.93393 ± 0.00288
NSGA-II	9	2364.03 ± 83.44	0.93398 ± 0.00233
NSGA-II	10	2357.86 ± 97.26	0.93415 ± 0.00272
NSGA-II	16	2370.66 ± 68.32	0.93379 ± 0.00191
All features	65	2556.2 ± 128.95	0.92861 ± 0.0036

Table 3 shows that removing features with zero average SHAP values from the 16 optimally selected features does not reduce model accuracy, though it does lead to increased variance in performance across experiments. This outcome provides stronger evidence that SHAP-based feature selection effectively maintains accuracy while refining the feature set.

4. CONCLUSIONS

Our study reveals that SHAP does not reliably select small subsets of informative features from whole dataset as it does not consistently rank features by their predictive value. SHAP and XGBoost also tend to assign high significance to features that contain the same information, indicating that these methods do not effectively eliminate redundancy. Our study shows that features selected by XGBoost generally outperform those selected by SHAP, consistent with findings from another study [11]. Interestingly, the study in [11]

reports that XGBoost's efficacy diminishes as feature count increases. Similarly, our study found SHAP to be more effective in refining smaller feature sets, maintaining model accuracy by eliminating non-contributive features.

However, SHAP requires more computation time and often selects features comparable or inferior to those selected by XGBoost. While SHAP and XAI are valuable for creating interpretable models, we recommend against relying solely on SHAP for feature selection. A hybrid approach that combines multiple methods could offer a more effective solution.

REFERENCE

[1] I. Guyon and A. M. De, “An Introduction to Variable and Feature Selection André Elisseeff,” 2003.

[2] S. Salcedo-Sanz, L. Cornejo-Bueno, L. Prieto, D. Paredes, and R. García-Herrera, “Feature selection in machine learning prediction systems for renewable energy applications,” Renewable and Sustainable Energy Reviews, vol. 90, pp. 728–741, 2018

[3] W. E. Marcílio and D. M. Eler, “From Explanations to Feature Selection: Assessing SHAP Values as a Feature Selection Mechanism,” in 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2020, pp. 340–347.

[4] Y. Liu, Z. Liu, X. Luo, and H. Zhao, “Diagnosis of Parkinson’s disease based on SHAP value feature selection,” Biocybern Biomed Eng, vol. 42, no. 3, pp. 856–869, Jul. 2022

[5] C. van Zyl, X. Ye, and R. Naidoo, “Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP,” Appl Energy, vol. 353, p. 122079, 2024

[6] S. Lundberg and others, “SHAP (SHapley Additive exPlanations),” 2023.

[7] A. Meepaganithage, S. Rath, M. Nicolescu, M. Nicolescu, and S. Sengupta, “Feature Selection Using the Advanced Shapley Value,” in 2024 IEEE 14th Annual Computing and Communication Workshop and Conference, CCWC 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 207–213

[8] A. Gramegna and P. Giudici, “Shapley Feature Selection,” FinTech, vol. 1, no. 1, pp. 72–80, Feb. 2022

[9] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II,” 2002.

[10] F.-A. Fortin, U. Marc-André Gardner, M. Parizeau, and C. Gagné, “DEAP: Evolutionary Algorithms Made Easy François-Michel De Rainville,” 2012

[11] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, “Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods,” J Big Data, vol. 11, no. 1, Dec. 2024